Abstract

Machine learning (ML) is widely applied in robotic technology. From facial recognition to nature language processing, it is becoming an integral part of life. Although ML was originally used in medicine over 50 years ago, only recently it has become highly popular thanks to the exponential advances in computational hardware witnessed in the past decade. An important application of ML in medicine is in the analysis of laboratory data for breast cancer diagnosis. Breast cancer is the most frequently diagnosed cancer among women. Currently, breast imaging for the detection and characterization of suspicious breast lesions relies on mammography and ultrasound. Mammography is the technique of choice for early detection of breast cancer. Although it is very sensitive at finding cancer, mammography results in many false positives. 20% of biopsied cases actually revealing cancer. The remainders turn out to be benign cases. The false diagnosis leads to potentially dangerous and unnecessary surgical procedure for otherwise healthy patients. Preventing benign biopsies is the most important way to improve the efficacy of mammography, especially as screening is becoming more and more widespread. ML methods for tumor classification have emerged as a proposed solution. In the medical industry, support vector machines (SVMs) and artificial neural networks (ANNs) are most common. In this work, SVMS and ANNs are critically analyzed. Each method is tasked with classifying various breast tumors as malignant or benign from a publicly available UCI Wisconsin Breast Cancer Dataset. After allowing the methods to train sufficiently, each of them undergoes an evaluation step towards establishing which of the methods is most suitable for breast cancer diagnosis.

Rationale

New and improved medical technologies are being invented every day to face some of the toughest diseases in the world. Breast cancer is the most frequently diagnosed cancer among women. Worldwide, breast cancer in females accounts for 23% (1.38 million) of the total new cancer cases and 14% (458,400) of the total cancer deaths in 2008 [1]. Currently, breast imaging for the detection and characterization of suspicious breast lesions relies on mammography and ultrasound. However, mammography suffers from major problems and only 20% of biopsied cases actually reveal cancer [2]. The remainders are all benign cases, and patients with such conditions are more often wrongly diagnosed to suffer from cancer based on false positives obtained from mammographic procedures. In such cases, such patients end up undergoing potentially dangerous and unnecessary surgical procedure. Preventing benign biopsies is therefore a critical need at this time, especially as screening by mammography has become most common. This clearly demonstrates a need for efficient breast cancer detection and diagnosis techniques, and ML methods for tumor classification are viewed as a viable solution.

Efficient machine learning models provide accurate diagnosis while being completely noninvasive. These models utilize existing available information, such as mammogram, ultrasound, and patient history data. Currently, Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) are the two most common machine learning methods in medicine. ANNs are a collection of mathematical models that emulate some of the properties of biological nervous system. The key element of the ANN paradigm is the novel structure of the information processing system. It is composed of many highly interconnected processing elements that are comparable to synapses in mammalian brains. Computer based diagnostic systems hold promising means against the challenges of clinical situations. A growing trend is noted the last decade in the use of other supervised learning techniques, namely SVMs, towards cancer prediction and prognosis [3]. SVMs use an efficient algorithm to statistically optimize a classification algorithm. While SVMs may sound much simpler, they are a powerful new technology. SVMs primary advantage is how little computing power is required to create an effective algorithm.

Both ANN and SVM methods of classification algorithms are extensively used in a wide range of problems posed in cancer research. However, SVMs have become more popular in the larger medical field. This raises an important question: which ML method, between the two, performs best in cancer tumor classification? A careful consideration of the literature in this area of research appears to suggest that **SVM will significantly outperform ANN in this specific cancer classification task.** However, a targeted study is necessary to weigh the advantages and shortcomings of the methods, and this project explores this question in detail.

Background

Although even most laymen understand the importance on medical imaging, many may not have in in-depth understanding of machine learning and how it can be applied to medical data analysis. One broadly accepted definition of machine learning is, "*If a machine learning algorithm is applied to a set of data, and some knowledge about these data, then the algorithm system can learn from the training data and apply what is has learned to make a prediction*" [4]. Machine learning is being applied in many areas outside of medicine, such as speech recognition, translation between languages, and autonomous navigation of vehicles. Some of these tasks were not previously feasible. Recent advancements in machine learning have made them possible. In the past, machine learning required enormous computational overhead, specific and structured data, and frequently could not successfully learn if any single point of data was missing. Newer algorithms are significantly more efficient and can gracefully accommodate omissions in data, and in some cases, can purposely remove data during the learning phase to make the algorithm more robust.

Machine learning uses many concepts that are unheard of in other sciences. The following list of key terms may aid in understanding how machine learning works [4].

Classification: Assigning of a class or label to a set of data. For instance, a machine learning model may mark some specific subset of data as "abnormal". The classifier might then try to determine whether this piece of data is significant enough to cause malignancy. [5].

Model: The set of weights or decision points learned by a machine learning system. Once learned, the model can be assigned to an unknown situation to predict which class that situation belongs to [5].

Algorithm: Series of steps taken to create the model that will be used to most accurately predict classes from the features of the training data [5].

Labeled Data: The set of examples (e.g. images), each with a correct "answer" (classification). For some tasks this might be the correct boundary of the tumor, and in other cases, it might be whether cancer is present or the type of cancer the lesion represents [6].

Training: The phase during which algorithms generate model with given labeled data. The set of weights or decision points for the model is updated until so substantial improvement is performance is achieved [6].

Testing: In some cases, a third set of examples is used for real-world testing. The algorithm system iterates to improve performance with a second set of data. Good performance with an unseen testing set can increase confidence that the algorithm will yield generalized and accurate results in the real world [6].

Node (Neuron): A part of a neural network that involves two or more inputs an activation function. The activation function typically sums the inputs and then uses some type of function and threshold to produce an output. Some activation functions include sigmoid, tanh, and rectified linear [7].

Method

Step 1: Data acquisition

A UCI Wisconsin dataset (1995) will be downloaded from the UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).

To train and evaluate a machine learning model, a sufficiently large dataset of mammogram samples must be acquired. Within this dataset, features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. This dataset contains 569 samples, each with 32 attributes.

Step 2: Data visualization

Create a correlation map.

Determining relationships within the data can aid in deciding which machine learning methods to use. Therefore, the relationships between their various attributes are visualized.

Step 3: Data pre-processing

Restructure the data and prepare for inputting into machine learning models.

This dataset is provided in a CSV file format. To accelerate the training process, all 569 samples are first loaded into working memory. The first attribute of each sample is an ID number and is discarded. Then the remaining 31 features are split:

Y (attribute 31): This final attribute is the tumor classification, malignant and benign, represented numerically as either 1 or 0. This feature is separated from the rest of the data. This is referred to as the *target* feature.

X (attributes 1 – 30): These remaining features are the *predictors* (mean radius, mean texture, mean perimeter, mean area, mean smoothness, etc.), which will be inputted into machine learning models.

This pre-processing is a necessary to step to ensure data compatibility with industry standard machine learning tools, such as TensorFlow and Scikit-learn. The data is then normalized, slightly tweaking outliers to fit a more general distribution. Both X and Y are then split into two smaller subsets. 80% of the data, 455 samples, is moved to a training set, and the remaining 20%, 114, samples are moved to a testing set. The training set will be directly inputted into both machine learning models during their respective training phases. The remaining testing data will be used to evaluate performance post training.

Step 4: Create an Artificial Neural Network.

Initialize the structure of an artificial neural network.

Modern machine learning frameworks greatly simplify model creation and evaluation. The most recommended machine learning frameworks for artificial neural networks are Tensorflow and Keras and are therefore preferred.

Generate a neural network with the following structure:

Layer 1 (Input) – 30 nodes, sigmoidal activation function.

Layer 2 – 4 (Hidden) – 512 nodes, rectified linear activation function.

Layer 5 (Output) – 1 node, sigmoidal activation function.

Various other neural network sizes and activation functions are acceptable and can be adjusted for optimal performance.

Step 5: Train the Artificial Neural Network.

Input training data and train the network until loss becomes acceptably low.

To learn, ANNs undergo several phases. The first phase, forward propagation, occurs when the network is exposed to the training data. The data is passed through the entire network and a prediction of whether a given sample is malignant or benign is generated. Initially, this prediction is essentially random. Next, a loss function estimates how well the model performed, in this case how accurately a sample was classified. Once the loss has been calculated, the information is propagated backwards, hence the name: backpropagation. Starting from the output layers of the network, the algorithm works backward, slightly adjusting edges and biases based on the loss. This training cycle is repeated until the model accuracy is sufficiently high or the model loss is sufficiently low.

Step 6: Create a Support Vector Machine

Initialize the structure of a support vector machine.

Modern machine learning frameworks greatly simplify model creation and evaluation. The most recommended machine learning framework for support vector machines is Scikit-learn and therefore preferred. Unlike artificial neural networks, there are no predisposed hyperparameters to tune.

Step 7: Train the Support Vector Machine

Input training data into the support vector machine and train until loss becomes acceptably low.

Unlike most other classifiers, SVMs only focus on samples that are most difficult to classify correctly. Other classifiers use all the data to evaluate relationships within the data. The rationale behind a support vector machine is; *if a model is sufficiently able to classify the most challenging samples, then it will be able to classify all other samples with an even higher degree of accuracy.* The SVM model is explicitly searching for the best separating line between the two classes of data. This is done by first searching for the two closest overlapping samples and finding a line, typically linear, that connects them. The SVM then declares that the best separating line is the line that bisects – is perpendicular to – the connecting line. This is repeated with many overlapping samples until the number of samples misclassified is minimized or, more generally, until the distance between the separating line and both classes of data is maximized.
Step 8: Evaluate both methods.

Use the testing dataset prepared in Step 3 to evaluate network performance.

During the data pre-processing step, the testing set was created to be intentionally isolated from network training. Post training, this data into inputted into each machine learning method to evaluate performance. A significantly lower performance in this testing data indicates a large amount of overfitting has occurred. This secondary evaluation is essential to ensure a machine learning models performance in a more generalized and real-world environment. These results can then be used to generate heatmaps to easily describe, true positives, true negatives, false positives, and false negatives.

References

[1] Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., & Forman, D. (2011). Global cancer statistics. *CA: a cancer journal for clinicians*, *61*(2), 69-90.

[2] Chen, H.-L., Yang, B., Liu, J., & Liu, D.-Y. (2011). A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. *Expert Systems with Applications*, *38*(7), 9014–9022. doi: 10.1016/j.eswa.2011.01.120

[3] Jalalian, A., Mashohor, S. B., Mahmud, H. R., Saripan, M. I. B., Ramli, A. R. B., & Karasfi, B. (2013). Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review. *Clinical Imaging*, *37*(3), 420–426. doi: 10.1016/j.clinimag.2012.09.024

[4] Sawarkar, S. D., Ghatol, A. A., & Pande, A. P. (2006). Neural Network Aided Breast Cancer Detection and Diagnosis Using Support Vector Machines. *Proceedings of the 7th WSEAS International Conference on Neural Networks*, 158–163. Retrieved from https://pdfs.semanticscholar.org/3df0/ec7c3feea8ec6a40bfae01d98b745ccb174c.pdf

[5] Ganesan, N., Venkatesh, K., Rama, M. A., & Palani, A. M. (2010). Application of neural networks in diagnosing cancer disease using demographic data. *International Journal of Computer Applications*, *1*(26), 76-85.

[6] Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.

[7] Agatonovic-Kustrin, S., & Beresford, R. (2000, April 13). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. Retrieved from https://www.sciencedirect.com/science/article/pii/S0731708599002721